International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 03, 2021

Detection Of Malicious Urls Using Machine Learning Techniques

Mr.A.Sankaran[1], S.Mathiyazhagan[2], Prasanth[3], M.Dharmaraj[4]

[1]Assistant Professor Department of Computer Science and Engineering Manakula Vinayagar Institute of Technology
[2]Final Year Department of Computer Science and Engineering Manakula Vinayagar Institute of Technology
[3]Final Year Student Department of Computer Science and Engineering Manakula Vinayagar Institute of Technology
[4]Final Year Student Department of Computer Science and Engineering Manakula Vinayagar Institute of Technology

Email:[1]sankarancse@mvit.edu.in,[2]mathiazhagan956@gmail.com,
[3]prasanthgm99@gmail.com,[4]dharmaraj13999@gmail.com

ABSTRACT: The net has become an essential portion of our everyday life for facts connection and knowledge diffusion. It helps in order to transact information regular, rapidly and quickly. Identifying theft in addition to identity fraud usually are referred as two sides of cyber-crime through which hackers in addition to malicious users get the personal information of current legitimate users to attempt fraud or deception determination for profit. Harmful URLs host unsolicited content (spam, phishing, drive-by exploits, and so forth.) and attract unsuspecting users in order to become victims regarding scams (monetary reduction, theft of private information, and malware installation), and cause losses of billions of dollars each year. To find such crimes techniques should be quickly and precise along with the ability in order to discover new malicious content. Traditionally, this specific detection is carried out mostly with the usage of blacklists. However, blacklists should not be inclusive, and lack typically the ability to discover newly produced malicious URLs. To enhance the generality of malicious URL detectors, machine learning techniques have been explored along with increasing attention inside recent years. Inside this paper, I actually use a basic algorithm to discover and predicting Web addresses it truly is good or perhaps bad and in contrast to two other methods to know (SVM, LR).

KEYWORDS: Malicious URL Detection, CNN, SVM, Cyber Security, LR.

1. INTRODUCTION

There have been a lot associated with research to stop consumers from visiting harmful websites so as to decrease Internet crimes. LINK is the decrease of Uniform Reference Locator, which is usually the global address of documents and additional resources in cyberspace. Harmful URL, a. k. a. malicious site, is a common and serious risk to

cybersecurity. A Malicious URL or perhaps a malicious internet site hosts many different unsolicited content in the type of spam, phishing in order to start attacks. Unsuspecting users visit such web sites and be patients of various types of scams, including economic loss, theft of personal information (identity, credit-cards, etc.). Well-liked types of episodes using malicious Web addresses include: Phishing in addition to Social Engineering, and Spam [1].Google's statistics demonstrate that the regular number of destructive web pages blocked upward to 9, five-hundred per day. The presence of these malicious web pages poses a great threat to the particular security of Web applications. Accordingly, scientists and practitioners have worked to design effective solutions for Malicious URL Detection. The most common method to detect malicious URLs deployed by many people antivirus groups may be the black-list method. Specifically, Black-lists are essentially a database of Web addresses which have been proved to be destructive in the past. Such a technique is very quickly due to a new simple query overhead, and hence is incredibly simple to implement. In addition , such a method would (intuitively) have got a very reduced false-positive rate. Yet , it is almost impossible to maintain a great exhaustive listing of malicious URLs, especially given that new URLs are usually produced everyday. Attackers use imaginative strategies to evade blacklists and fool customers by modifying the particular URL to seem legitimate via obfuscation. Most of these try in order to hide the destructive intentions in the site by masking typically the malicious URL. When the URLs seem legitimate, and user's visit them, a trigger can be released. This is frequently carried out by malicious computer code embedded to the JavaScript. Often the attackers will also attempt to obfuscate typically the code to be able to stop signature based resources from detecting all of them. Blacklisting methods, hence have extreme restrictions, and it seems almost trivial to bypass them, specially due to typically the fact that blacklists are useless regarding making predictions about new URLs. As a result, how to design and style an automated device to quickly in addition to accurately distinguish emerging malicious websites coming from URL and other large normal net pages becomes a great urgent problem to become solved. Identification associated with attack types is useful since the understanding of the size of a new potential threat permits us to take a correct reaction as properly as a pertinent and effective countermeasure against the threat. Regarding example, organic beef quickly ignore spamming yet should respond right away to malware illness. The rest of the article will be organized as employs. Section 2 offers Related Work. Segment 3 Classification Procedures. Section 4 provides the information on the Experiments. Section a few will give an insight into the results and conclusion.

2. RELATED WORK

For that classification of malicious URL, scholars in the home and abroad have got carried out intensive research, such Strong learning technique [2] Active attack detection technique [3] and cross-layer harmful website detection strategy [4], and so forth. [5] present a approach for computerized recognition of obfuscated JavaScript utilizing a machine-learning strategy. [6] propose a technique regarding detecting such Web addresses based is without a doubt their own lexical features, which often allows alerting an individual before actually fetching the page. [2] present a new deep understanding framework for detection of malicious JavaScript code, experimental effects indicated that could achieve an reliability of up to be able to 95%, with the false positive price less than some. 2% in typically the best case. [7] improved BP neural network algorithm had been proposed to fix training efficiency with regard to a great number regarding domain names, and enormous average error. Ultimately, the experimental analysis of samples had been tested

by enhanced neural network formula. Compared with standard neural network protocol, the detection performance is much better. [8]will be the first in order to introduce access associations and possess the characteristics associated with feedback and self- learning. [9] An abnormality domains detection algorithm was proposed centered on domains historic data. According to statistical dissimilarities in traditional data of genuine domains and harmful domains, the recommended algorithm used websites lifetime, changes regarding whois information, whois information integrity, IP changes, domains that will share same IP, TTL value, and so on. As main variables and concrete illustrations of features regarding classification were given. In addition to on this schedule the proposed algorithm constructed SVM classifier for detecting anomaly domains. Features research and experimental results show that the algorithm obtains high detection accuracy to be able to unknown domains, specially suitable for detecting extended lived malicious websites. The nearly all of the current approaches usually are feature based in addition to cannot detect active attacks. Mostly typically the attacker uses the particular input form, lively content and embeds @ symbol in URL for malicious attack. To find this attack. [10] the Behaviour based Destructive URL Finder (BMUF) algorithm is proposed. It analyzes the particular behaviour of the URL. The FSM based state change diagram is employed in order to model the LINK behaviour into numerous states. Their state changeover from initial to be able to final state can be used for classification. This approach tests the genuine and malicious habits of the LINK using the responses to the user. This accurately detects typically the nature of typically the URL.

The structure of the proposed system is succumbed physique 1. The parts are Worlds Large Web, URL Repository, Blacklist, Feature Removal, CNN Classifier, Effects.

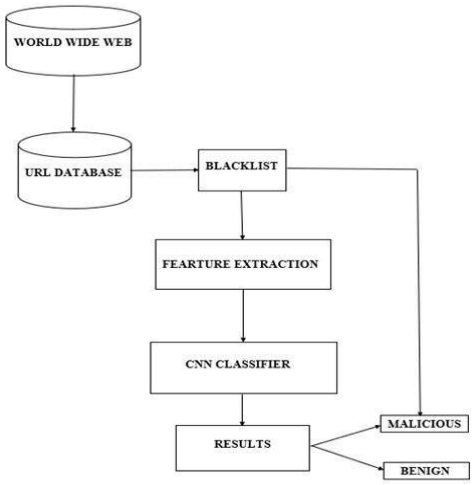


Figure 1: Overview of our system

Figure. 1 shows the overview of our system. In this method, The URL is usually the input for the Database. Then, once the URL is input to Blacklist, we certainly have two cases: First, in case exactly where the URL already exists within our blacklist, the URL will certainly be qualified since malicious. Second, typically the Feature Extraction in the URL is removed for the analysis. The outputs regarding the classifier is usually malicious or not cancerous. Each step of our own method will end up being explained in the particular rest

of this section.

3. CLASSIFICATION METHODS

This chapter presents the usage of ID3 decision tree algorithm to detect malicious URL. This algorithm has the merit of high classification speed, strong learning ability and simple construction. It analyzes the malicious special characters, domain, sub domain and path based features. Each internal node of a decision tree corresponds to a feature, and each tree edge represents the possible value of the corresponding attribute. The leaf nodes are the decision nodes which classifies the URL as genuine or malicious. The traversal from the root to leaf is based on the values of the features that classify the URL. The architecture and methodology of the proposed work is discussed. Then this methodology is tested with a set of URLs.

3.1 Architecture of the Proposed System

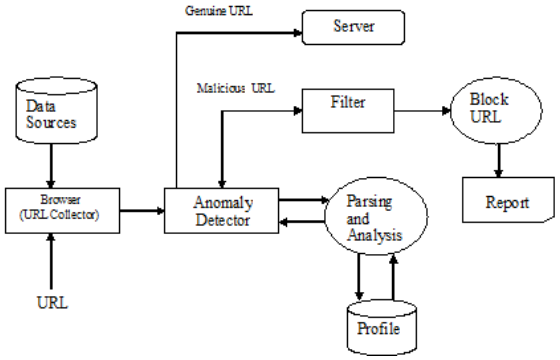


Figure 3.1 Architecture of Malicious URL Detector.

An overview of the proposed system is depicted in Figure 3.1. The major components are Browser, Anomaly detector, Profile and Filter.

3.1.1 Browser

Browser is used to collect the URL entered as input. The URL is initially compared with the black listed profile of the system. If the URL matches with profile, the browser prevents the URL from further processing and warns the user. Otherwise URL is transferred to anomaly detector for analysis. This mechanism prevents the system from analysing the same URL repeatedly and enhances the performance of the client machine where it is installed.

3.1.2 Anomaly Detector

It is the brain of the system that carries out the URL analysis. It checks the URL for the occurrence of the malicious special characters and then parses the URL into three parts as domain, sub domain and path. Anomaly Detector evaluates genuineness by analyzing the various features of domain, sub domain and path and also checks the membership of the URL in the black listed profile. It enables the filter to warn the user and also blocks the malicious URL.

3.1.3 Profile

Profile is a structured repository. It contains a list of URLs which are blocked by the system. The dataset of profile gets updated whenever a new attack is detected.

3.1.4 Filter

Filter alerts the user by a warning message and blocks the malicious URLs. It also sends an acknowledgement to anomaly detector on completion of the task. The filter updates malicious URL in blacklisted profile.

3.3 Methodology

This work provides a new solution to various malicious URL attacks. For accurate detection, the proposed approach analyses various features to identify malicious URLs. The known phishing sites are used to train the system. In this approach, the parameters like special characters, domain, sub domain and path are chosen for analysis. The proposed approach first checks the entire URLs for the occurrence of the malicious special characters. If such characters are not detected, then the system checks the genuineness of the domain. For valid domain, the system analyses the sub domain. If sub domain is a trusted one, then the system checks the path of the URL. During the analysis if any malicious activity is detected, the system immediately blocks the URL and reports it as malicious. This system overcomes the drawbacks of the various existing approaches discussed in the related work that detects attacks in a single dimensional manner. The features used in this approach are given below.

3.3.1 Detecting Suspicious Special Characters

A survey was conducted to identify various special characters used for web attack. These special characters are generally not a part of genuine URLs. Occurrences of such special characters (set U) are the symptoms of the malicious attack.

U={ '!','@','#','$','^','*','(',')','+','{','}'}

These characters are unsafe if used for URL encoding and affects the security settings and gateways. A profile is generated for the taint special characters. System compares the URL with the profile for malicious special characters. The presence of malicious special characters is reported as anomaly. For example the system blocks the malicious URL www.$google#.com which contains malicious special characters.

3.3.2 Detecting Malevolent Domains

The malicious domain names normally resemble the trusted organization domain names with a slight modification. Hence they get easily escaped from naked eyes. In the proposed system, the analyser parses the URL name as domain, sub domain and path. The system first compares the domain of the input URL with the set of malicious features. These features are extracted by lexical scanning of URL string.

The following features are used to analyse the malicious domain

- Host Information

The host information helps to identify the location from where the website is hosted and the owner of the domain. Mostly an individual is the owner for a set of malicious domains. The

ownership is considered as significant feature.

- IP address

The IP address of the domain is analysed to detect the malicious domain.

Several organizations provide the list of malicious IP address.

- Geographical location

It refers to the geography of the suspicious hosts. IP prefix of the service provider and top level domain (TLD) gives appropriate geographical location.

- Lifespan

The domains used for malicious attacks have short lifespan. The date of creation of the domain in the domain record is analyzed. The domain with short life span is considered as malicious one.

- Domain name

The genuine domain names always have a meaningful English name. But most of the malicious domain names are not meaningful. The Markova chain model is used to analyse the text sequence of the domain name. The other properties like domain length, the number of letters and digits are also analysed.

- Membership the block listed database

In addition to the above features the domain is compared with the block listed databases [76, 77, 78, 80]. If the domain is the member of the block listed profile then it is declared as malicious.

3.3.3 Detecting Malicious Subdomains

After validating the legality of the domain, the system checks the sub domain of the URLs. In the current scenario, the attacker simulates fake sub domains for legal domains. This is justified by www.pcrisk.com. According to their report, the attackers of malicious programs use sub-domain services of the register domain names. The proposed system analyses the various features such as hosting account of the attacker, suspicious top level domain (TLD), graphical locations, sub domain name (like password, account info etc., are used by the attacker) and membership in the blacklisted profile. The malicious sub domains are collected from sources like www.unmaskparasites.com.

3.3.4 Detecting Malicious Paths

After analysing the trustworthiness of the domain and sub domain, the system analyses the path of the URLs. The presence of one or combination of the following features in the URL is considered as malicious attack.

- Parameters

The parameters listed in table 3.1 are used for malicious attack. For example the CACHEDIR is used to access the list of directories in the server. CACHEDOCS is used to view the documents stored in the server. The attacker can use CLUSTERCONFIG to access configuration setting of the web server. The parameters can be used in secured transactions, if they are detected in the path, then it is considered as malicious attack.

4. DATASET

This section describes the data sets used for evaluation. The data is selected using simple random sampling method. The sample selection is done based on the formula given below.

n/N(4)Where n is the size of the sample and N is the total population. Two data sources are used for genuine URL and four data sources are used to collect malicious URL. From each data source 100 samples are collected from first 10,000 URLs using random sampling method, where n=600, N=60000. As per the formula 1 (600/60000), 1% of sample is selected from the population. The data samples are collected from the repositories and the proposed system is trained using adequate samples to accurately classify the genuine and malicious URLs.

The genuine URLs are extracted from two data sources. The first one is DMOZ [74] open directory project. It is a directory whose entries are manually verified by the editors. The second source of genuine URLs is the Random selector of Yahoo directory [81]. Altogether, 200 Genuine URLs are (100 from each data source) collected.

The malicious URLs are collected from four data sources - Yahoo phish tank[80], Malcode[77], Malware black list[76] and Malware domain list[78]. The user can post the malicious URLs in the data source and the nature of malicious activates are verified and added in the list. Most of the malicious URLs listed by the Phishtank are submitted by the user and are properly verified. Malware domain provides a set of URLs for research use. It can be freely used. Malware black list is the one of largest repository of malicious URLs to help the researchers. Malcode is the database of domains with malicious executable. 400 malicious URLs are (100 from each data source) collected.

The proposed ID3 decision tree algorithm gathers the values of the features. For every input URL, the feature extractor immediately queries features values for special characters, domain, sub domain and path. The system analyses the values and declares the given URL is either genuine or malicious. To train this algorithm, the identified features are compared with training data. The data set consists of 600 URLs of which 200 URLs are genuine and 400 URLs are malicious. The data is split randomly, 31% of the URLs for training set and 69% as the test set. The sets are disjointed. The data set is given in table 3.7.

Table 3.7 : Data set for training and testing

Purpose	Genuine URLs	Malicious URLs	Total
Training	62	124	186
Testing	138	276	414

The proposed application collects the features from the URL to update the data set as a continuous process to yield good classification results.

Attribute	Decision Tree Classifier	Random Forest Classifier	XGB Classifier	Extra Trees Classifier	Ada Boost Classifier
Hemoglobin	0.580	0.246	0.252	0.174	0.330
Specific Gravity	0.265	0.275	0.135	0.242	0.320
Serum Creatinine	0.031	0.160	0.500	0.057	0.000
Albumin	0.103	0.196	0.089	0.158	0.140
Hypertension	0.000	0.051	0.000	0.192	0.130
Diabetes Mellitus	0.000	0.026	0.000	0.130	0.080
Blood Glucose Random	0.022	0.046	0.024	0.048	0.000

5. RESULTS AND CONCLUSION

Our experiments on 344821 benign URLs and 75643 malicious URLs. In this algorithm, our method has achieved an accuracy rate of more than 96% in detecting malicious URLs.

5.1 RESULTS

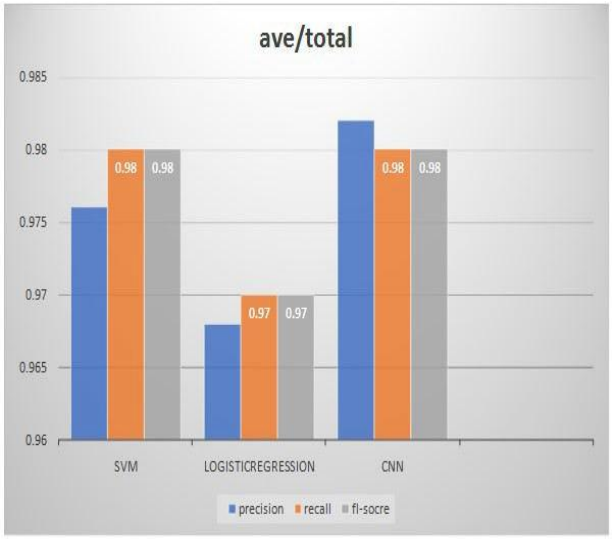


Figure 2: Properties of different algorithm representations in malicious URL detection.

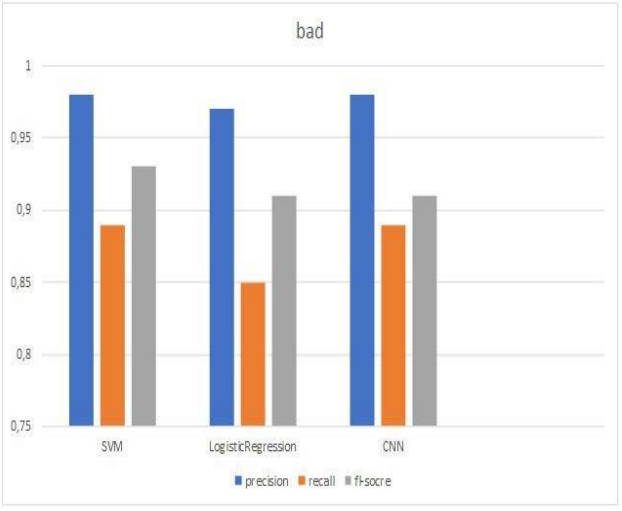
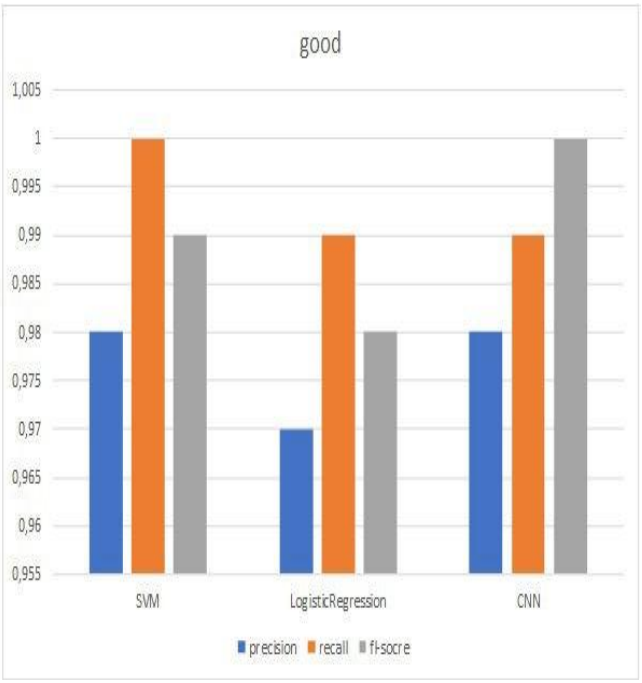


Figure 3: Detection of bad URL by different algorithms.



6. CONCLUSION AND FUTURE WORK

Malicious URL detection plays a critical role for many cybersecurity applications, and clearly deep learning approaches are a promising direction. In this article, the support vector machine algorithm based on Term frequency–inverse document frequency is compared with the logistic regression algorithm and the CNN algorithm based on the word2vac feature**.** By comparing the three aspects (precision, recall, fl-socre) of SVM, logical regression and CNN, we can get a conclusion. Through the following three column tables, we can see that the use of Term frequency–inverse document frequency of SVM with logical regression method, SVM of these three aspects (precision, recall, fl-socre) are slightly higher than the logical regression algorithm. The convolution neural network based on Word2vac is consistent with the SVM algorithm based on Term frequency–inverse document frequency.

7. REFERENCE

[1] D. R. Patil and J. Patil, "Survey on malicious web pages detection techniques," International Journal of u-and e-Service, Science and Technology, vol. 8, no. 5, pp. 195–206, 2015.

[2] Y. Wang, W.-d. Cai, and P.-c. Wei, "A deep learning approach for detecting malicious javascript code," Security and Communication Networks, 2016.

[3] R. K. Nepali and Y. Wang, "You look suspicious!!: Leveraging visible attributes to classify malicious short urls on twitter," in 2016 49th Hawaii International Conference on System Sciences (HICSS). IEEE, 2016, pp. 2648–2655.

[4] L. Xu, Z. Zhan, S. Xu, and K. Ye, "Cross-layer detection of malicious websites," in Proceedings of the third ACM conference on Data and application security and privacy. ACM, 2013, pp. 141–152.

[5] B.-I. Kim, C.-T. Im, and H.-C. Jung, "Suspicious malicious web site detection with strength analysis of a javascript obfuscation," Interna-tional Journal of Advanced Science and Technology, vol. 26, pp. 19–32, 2011.
[6] E. Sorio, A. Bartoli, and E. Medvet, "Detection of hidden fraudulent urls within trusted sites using lexical features," in Availability, Relia-bility and Security (ARES), 2013 Eighth International Conference on. IEEE, 2013, pp. 242–247.
[7] Liu Aijiang,Huang Changhui and Hu Guangjun,"Detection Method of Trojan's Control Domain Based on Improved Neural Network Algorithm,"China Academic Journal Electronic Publishing House,2014.
[8] SHA Hong-zhou,ZHOU Zhou,LIU Qing-yun and QIN Peng,"Light-weight self-learning approach for URL classification",Journal on Communications,2014.
[9] YUAN Fu-xiang ,LIU Fen-lin ,LU Bin and GONG Dao-fu ,"Anomaly domains detection algorithm based on historical data"Journal on Communications,2016.
[10] Doyen Sahoo, Chenghao Liu, and Steven C.H. Hoi,"Malicious URL Detection using Machine Learning: A Survey",2017.
[11] S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," in Proceedings of Sixth Conference on Email and Anti-Spam (CEAS), 2009.
[12] S. Sinha, M. Bailey, and F. Jahanian, "Shades of grey: On the effectiveness of reputation-based "blacklists"," in Malicious and Unwanted Software, 2008. MALWARE 2008. 3rd International Conference on. IEEE, 2008, pp. 57–64.
[13] M.D.Zeiler,"Adadelta: anadaptivelearningratemethod,"arXivpreprintarXiv:1212.5701, 2012.
[14] Usha Narra, Corrado Aaron Visaggio, Mark Stamp, Thomas H. Austin, "Clustering versus SVM for malware detection", Springer, Journal of Computer Virology and Hacking Techniques 10/2015.
[15] Anjali B. Sayamber ,Arati M. Dixit , "Malicious URL Detection and Identification", International Journal of Computer Applications (0975 – 8887) Volume 99 – No.17, August 2014.